学校编码：10384　　分类号_____密级_____
学号：24320121152281　　UDC_____

厦门大学

硕 士 学 位 论 文

基于 BTM 的个性化推荐系统研究与应用

Research and Application of Personalized Recommender System Based on BTM

叶伟隆

指导教师姓名：王 备 战 教 授
专 业 名 称：计算机软件与理论

论文提交日期：2015 年 4 月
论文答辩日期：2015 年 5 月
学位授予日期：2015 年 6 月

指 导 教 师：_____

答辩委员会主席：_____

2015 年 6 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果,均在文中以适当方式明确标明,并符合法律规范和《厦门大学研究生学术活动规范（试行)》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名):

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

随着互联网的高速发展和数据的爆炸性增长，用户面临着日益严重的信息过载问题，社会化媒体的兴盛使用户更加容易淹没在信息的海洋中。推荐系统作为一种比搜索引擎更加高效的信息过滤技术，逐渐地成为各个社会化媒体的主要功能之一。

传统的推荐技术大多基于用户关系，难以有效地挖掘用户兴趣。本文在结合一元混合模型和 LDA 模型的基础上，引入了 BTM 用户兴趣建模技术，较好地解决了传统的向量空间模型维度高、矩阵稀疏和一词多义等问题，同时针对社会化媒体中的用户特征和数据特点进行了改进，并基于此设计和实现了一个包含数据收集模块、预处理模块、兴趣分析模块和个性化推荐模块的个性化推荐系统。论文的主要工作如下：

首先，研究和探讨了用户建模和推荐算法等推荐系统相关技术，其中重点研究了基于内容的推荐以及协同过滤推荐，并结合具体的应用场景分析了各自的优点和缺点。

其次，本文结合一元混合模型和 LDA 模型，引入了基于 BTM 的用户兴趣建模方法并详细阐述了 BTM 的原理和建模过程，同时针对社会化媒体中的用户特征和数据特点进行了改进，通过在真实数据集上的实验，验证了 BTM 在社会化媒体中用户兴趣建模的可行性和有效性。

最后，基于以上研究成果，本文设计并实现了一个包含数据收集模块、预处理模块、兴趣分析模块和个性化推荐模块的个性化推荐系统，并在真实数据集上进行模拟推荐，实验结果表明针对社会化媒体的 BTM 建模方法具有良好的效果，能有效地帮助用户发现其感兴趣的信息。

关键词：主题模型；推荐系统；文本挖掘

Abstract

With the rapid development of Internet and the explosive growth of data, users are facing an increasingly serious problem of information overload. In recent years, users are more likely to drown in the sea of Information because the rise of social media. Recommender system, as a more efficient information filtering technology better than Search Engine, is now becoming one of the main functions of every social media.

Most of the traditional recommendation technology is based on the relationship of users, and it's difficult to mine users' interest effectively. This dissertation introduces BTM modeling technology by combining Mixture of unigram and LDA model, and it’s improved according to the user behavior and data characteristic of social media. Based on these result, this paper designs and implements a personalized recommender system that includes data collecting module, preprocessing module, interest analyzing module and personalized recommending module. The main work of this dissertation is as follows:

Firstly, personalized recommender system technology is studied in this dissertation, including user modeling technology and recommending algorithms.

Secondly, this dissertation studies and discusses BTM modeling technology by combining Mixture of unigram and LDA model, and improved the modeling result according to the user behavior and data characteristic of social media. As the experiment result shows, this solution has a better effect.

Finally, a personalized recommender system includes data collecting module, preprocessing module, interest analyzing module and personalized recommending module is designed and implemented in this dissertation. Experiment result shows that this system has a good effect, it can help users find their interests effectively.

Keywords: Topic Model; Recommender System; Text Mining

目 录

Contents

第一章 绪论

1.1 研究背景与研究意义

随着互联网的高速发展和普及，信息呈现爆炸式增长，而近年来以社交网络为代表的社会化媒体的兴盛，使用户更加容易淹没在信息的海洋中。据官方财报显示，截至 2014 年第四季度，Facebook 每月活跃用户数达到 13.9 亿，其数据仓库每日新增的数据量超过 600TB[1]，而国内的新浪微博每月活跃用户数也有 1.6 亿，所有用户每天发布的微博数量超过 1 亿条，图 1-1 为 2014 年国内社会化媒体格局图。

图 1-1 2014 年国内社会化媒体格局

社会化媒体带来的海量信息已经远远超出了用户的接收和消化能力，针对这种信息爆炸带来的信息过载问题[2]，搜索引擎和推荐系统正逐渐取代传统的门户网站，成为更加高效的信息过滤技术。同时，数据的增长和增值也带来了互联网形态的改变，出于隐私和数据保护等方面的考虑，以社交网络为代表的社会化媒体正逐渐地从自由和开放走向封闭和私有，从根本上动摇了搜索引擎赖以生存的

生态基础。

与搜索引擎被动地接受用户输入不同，推荐系统通过收集和处理用户数据，主动地向用户推荐符合其兴趣的内容，更加符合社会化媒体以用户为中心的特性。目前大多数社会化媒体都应用了推荐系统相关技术：Facebook、Twitter 和新浪微博等社交网络平台上都出现了个性化的用户推荐、内容推荐甚至是广告推荐；豆瓣网可以为用户推荐其可能感兴趣的书籍、音乐和电影；Yelp 和大众点评等网站为用户推荐美食；Youtube 和优酷等网站为用户推荐视频……而这些应用的推荐效果参差不齐，尤其是社交网站上推荐的无关信息很大程度上降低了用户体验。因此，如何有效地帮助用户筛选出感兴趣的信息，不仅可以改善用户体验，提高用户活跃度，还有利于提升平台竞争力，提供新闻、游戏、广告等社会化媒体个性化服务。

个性化服务的核心是匹配用户的需求或用户兴趣[3]，即用户兴趣模型。随着智能设备和移动互联网的发展和普及，以及线上到线下（Online to Offline，O2O）商业模式的出现，社会化媒体越来越深入到我们的日常生活中，随之带来更加真实、更加多元、更加丰富的海量用户数据，如何针对社会化媒体建立合适的用户兴趣模型，推荐其可能感兴趣的内容，具有重要的理论和实践意义。

1.2 研究现状与存在问题

推荐系统的概念在上个世纪九十年代中期被提出[4]，经过二十多年的发展，由最开始关注的电子商务领域，逐渐扩展到包括社会化媒体在内的其他应用领域。受到电子商务网站和电影网站的影响，早期的个性化推荐主要解决的是评分预测问题，即预测用户对某个项目的评分。然而，由于用户显式数据难以获得，往往造成评分矩阵太过稀疏，难以完整地捕捉用户的兴趣。因此，近年来对个性化推荐的研究更注重解决 Top-k 推荐问题，即如何为用户生成个性化推荐列表。

目前主要的推荐算法有基于内容推荐和协同过滤推荐，基于内容的推荐对用户的历史项目特征和属性进行分析和学习，寻找相似度最高的项目进行推荐；协同过滤推荐则通过计算用户之间的相似度，基于相似用户具有相近兴趣的基本思想进行推荐。随着研究的深入以及推荐系统在商业领域的广泛应用，存在的一些

问题也逐渐出现，目前个性化推荐系统主要面临着以下挑战：

1、冷启动问题

推荐系统需要根据用户的历史行为预测其兴趣偏好，冷启动问题即历史数据为空的新对象难以获得推荐，主要分为用户冷启动、项目冷启动和系统冷启动问题。目前已有研究提出不同的解决方案[5]：针对用户冷启动问题，可利用用户的注册信息，要求用户主动填写相关兴趣信息，或者通过用户地理位置、社交网站等渠道获取用户信息；而针对项目冷启动问题，可利用项目内容信息的相似度进行过滤，如通过自然语言处理技术抽取关键词等。

2、矩阵稀疏问题

协同过滤推荐算法通过用户-项目评分矩阵进行推荐，由于在实际应用中该矩阵数据量大且异常稀疏，对其进行相似度计算往往花费较大，且推荐效果不理想。针对该问题有学者提出了不同的解决方案，如采用基于人口统计学的方法进行过滤、引入语义信息建立主题模型来降低维度等[6]。本文的研究内容也主要针对该问题。

3、兴趣漂移问题

由于用户的兴趣是变化的，具有长期兴趣和短期兴趣的特点，尤其在社会化媒体中，用户的行为不仅仅与自身兴趣有关，还与当下的社会背景、用户关系网络中的用户行为相关，因此对用户兴趣模型的更新和预测也是研究的一个热点和难点，有研究通过引入用户近期活动权值[7]来进行调整推荐。

用户兴趣模型的建立是个性化推荐系统的首要工作[8]，也是影响最终推荐效果的关键因素。目前的用户模型表示方法主要有向量空间模型（Vector Space Model，VSM）、主题模型（Topic Model）和神经网络模型等。向量空间模型由于其简单通用，已成为目前应用最成熟且最广泛的模型，而概率主题模型因为在高维度、数据稀疏等方面的优势，在提出之后不断地发展和完善。

由于社会化媒体中的数据文本较短、噪声较多、特征词少，采用传统的向量空间模型方法建立用户兴趣模型面临着数据维度高、数据稀疏和一词多义等问题，而主题模型通过非监督学习方式生成，无需事先对语料库进行特征抽取，更适合社社会化媒体中用户兴趣模型的建立。常用的主题模型如隐含狄利克雷分布

（Latent Dirichlet Allocation，LDA）模型[9]虽然可以有效降低维度，解决稀疏性问题，但也存在着短文本处理低效、用户兴趣动态变化等问题，还有待进一步的分析研究。

社会化媒体平台还具有平台封闭、数据内容丰富、信息传播快速、用户交互密切等特点，传统的个性化推荐技术不能很好的适用于社会化媒体平台，难以准确地描述和构建用户的行为特征和兴趣模型，也吸引了越来越多的关注和研究，例如 Chen 等人[10]通过比较多种用户数据来源，发现基于用户发布的内容构建的用户兴趣模型效果更好；Wu 等人[11]通过对用户发布的内容进行处理，应用 TF-IDF 统计方法和 PageRank 思想提取关键词以构建用户兴趣模型；Weng 等人[12]将用户发布的所有内容合成一个文档，并结合 PageRank 思想提出 TwitterRank 模型生成用户兴趣；陈文涛等人[13]认为概率主题模型要比向量空间模型更适合社交网络，同时通过对比三种较为常见的主题模型，证明 TwitterLDA[14]更适合预测新文档或新用户，AuthorLDA[15]可产生较高区分度的主题，而 UserLDA[12]和 AuthorLDA 能更好的刻画用户的社交关系。

1.3 论文主要工作

传统的文本挖掘方法不能很好的处理社会化媒体平台上的用户数据，以向量空间模型为代表的用户兴趣模型没有考虑语义信息，并且存在维度高和矩阵稀疏的难题，难以有效的进行个性化推荐。本文针对社会化媒体中的信息过载问题，引入了一种结合一元混合模型和 LDA 模型的词对主题模型（Biterm Topic Model，BTM）用户兴趣建模方法，同时针对社会化媒体中的用户特征和数据特点进行了改进，基于此设计并实现了一个包含数据收集模块、预处理模块、兴趣分析模块和个性化推荐模块的个性化推荐系统，最后通过在真实数据集上进行实验，结果表明该方法具有更好的性能。

论文的主要工作包括：

1、研究了个性化推荐系统的相关技术，包括用户建模技术和推荐算法技术，并结合具体应用场景分析了各个推荐算法的优缺点，详细介绍了一元混合模型和 LDA 模型等多种主题模型。

2、分析了社会化媒体中的用户特征和信息特点，结合一元混合模型和 LDA 模型引入了基于 BTM 的用户兴趣建模方法，同时针对社会化媒体中的用户特征和数据特点进行了改进，并通过真实数据集上的实验证明了该方法的可行性和有效性。

3、根据社会化媒体平台特性，基于 BTM 用户兴趣模型设计并实现了一个包含数据收集模块、预处理模块、兴趣分析模块和个性化推荐模块的个性化推荐系统。通过在新浪微博数据集上进行的推荐实验，表明该推荐系统具有良好的效果，可以很好地向用户推荐其可能感兴趣的信息。

1.4 论文结构安排

本文共为分六章，各章主要内容如下：

第一章 绪论。阐述了社会化媒体平台个性化推荐的研究背景和意义，研究现状及存在问题，并简述了论文的主要工作。

第二章 推荐系统相关技术研究。介绍了向量空间模型和主题模型等用户兴趣模型相关技术，并介绍了几种个性化推荐算法及算法评价标准，并结合具体应用场景分析了各个推荐算法的优缺点。

第三章 基于主题模型的用户兴趣研究。结合一元混合模型和 LDA 模型引入了基于 BTM 的用户兴趣建模方法，同时针对社会化媒体中的用户特征和数据特点进行了改进，并通过真实数据集上的实验证明了模型的可行性和有效性。

第四章 基于用户兴趣的推荐系统设计与实现。基于 BTM 用户兴趣模型设计并实现了一个包含数据收集模块、预处理模块、兴趣分析模块和个性化推荐模块的个性化推荐系统。

第五章 系统实验及结果分析。通过在新浪微博平台上进行的推荐实验，表明了该推荐系统具有良好的效果，可以很好地推荐用户感兴趣的信息。

第六章 总结与展望。总结了论文所做的工作，同时介绍了下一步的研究工作。